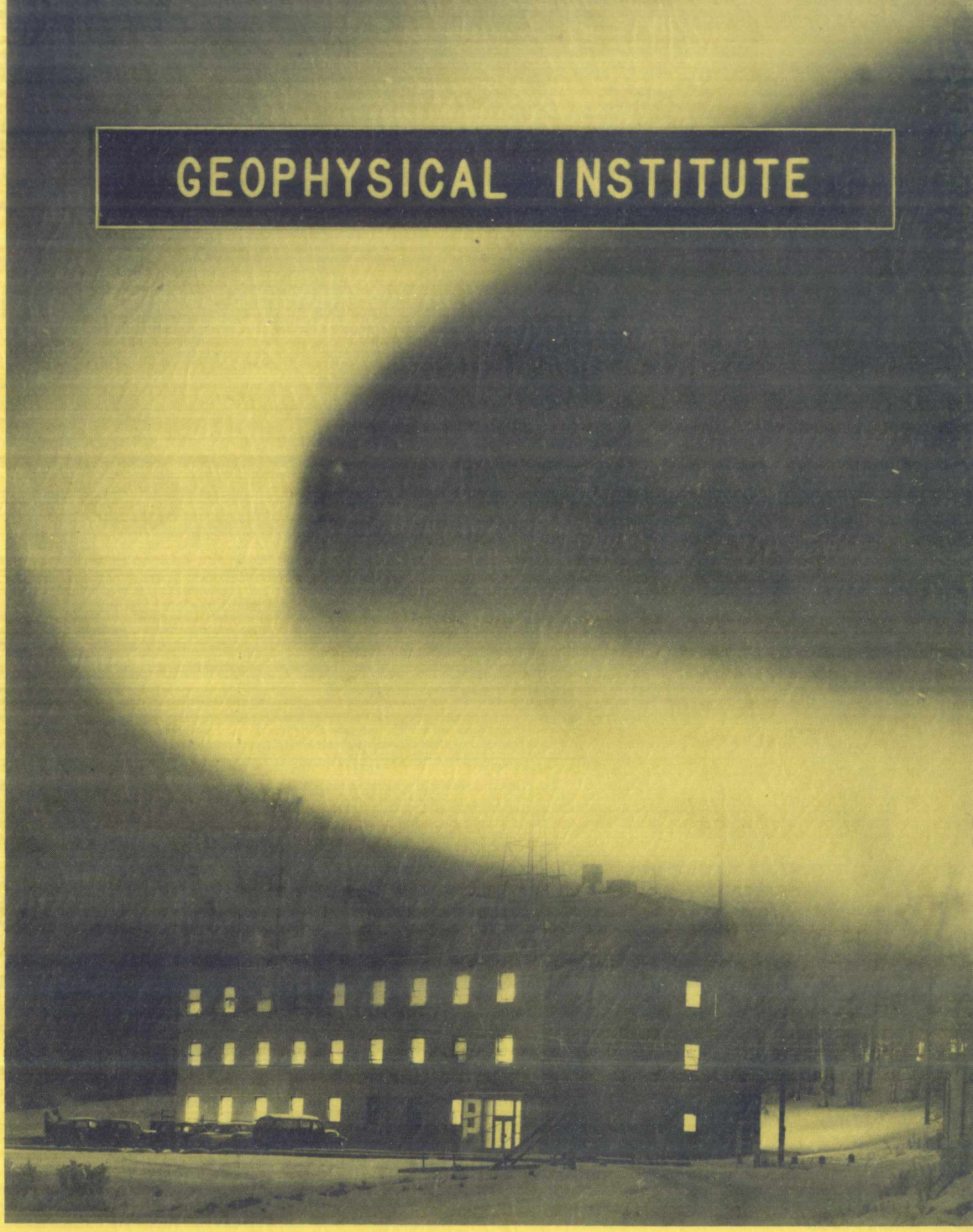


UNIVERSITY
OF ALASKA

COLLEGE
ALASKA



Geophysical Research Report No. 4

# A MAGNETO-IONIC THEORY OF THE AURORA

by

G. C. Reid

December 1958

Geophysical Research Report No. 4

GEOPHYSICAL INSTITUTE

AT THE

UNIVERSITY OF ALASKA

# A MAGNETO-IONIC THEORY OF THE AURORA

By

G. C. Reid

December 1958

C. T. Elvey, Director

# A MAGNETO-IONIC THEORY OF THE AURORA

G. C. Reid

## ABSTRACT

A qualitative description of the development of a typical auroral display as the result of an electrical discharge in the ionosphere is presented. The prime cause of the discharge is taken as the potential difference existing between points in the interplanetary medium as a result of an interaction between charged particles of solar origin and the earth's magnetic field. The characteristics of the occasional very intense aurorae visible over large areas of the earth are discussed, as well as the normal diurnal and seasonal variation of auroral occurrence. The origin of the electric field is discussed, and a possible explanation in terms of particles trapped in the earth's magnetic field is presented.

# A Magneto-Ionic Theory of the Aurora

## Introduction

For many years the cause of the aurora has been one of the outstanding problems of geophysics. Many theories have been advanced, all of which have been subjected to very serious criticism, and none of which can be taken as giving a complete picture of the auroral phenomenon, though many of them do present faithful pictures of aspects of this complicated process. It is realized that the ideas which will be presented in this paper will be subjected to similar criticism, but it is hoped that they may prove to be a step towards the final explanation of a beautiful and intriguing natural phenomenon.

No description or criticism of current auroral theories will be given here, and reference can be made to an excellent review article by Chamberlain[1], containing a very complete bibliography of auroral theory from the days of Aristotle.

Perhaps the major objection to most of the previous theories has been their lack of consideration of the observed facts of auroral morphology; many of the theories have been concerned with an explanation of the magnetic storm effects which accompany aurora, rather than with the optical phenomenon itself. Notable exceptions to this have included the theories of Störmer[2], Lebedinski[3], and, to some extent, of Alfven[4], as well as the treatment given by Chamberlain[5], of the luminosity-height characteristics of an auroral ray. The present paper will take the opposite view (probably equally reprehensible), and deal with the optical phenomena, saying very little about the associated magnetic effects.

The aurora is treated as an electrical discharge phenomenon in the earth's ionosphere, the prime cause being an electric field set up by the interaction between charged particles of solar origin and the earth's magnetic field. As such, the idea is by no means new. Martyn's[6] theory involves the production of a potential difference between the ionosphere and some point in interplanetary space of the order of $10^6$ volts, but this field is treated merely as a means of accelerating particles into the atmosphere, a process which cannot in itself explain the detailed picture of an auroral display. The idea of a discharge is a feature of Alfven's[4] theory, though the form of the discharge which will be considered in this paper is radically different from that described by Alfven.

Consequences of the Application of an Electric Field to the Ionosphere

A very simplified diagram of the initial conditions is shown in Fig. 1. A potential difference is assumed to exist between two points in the equatorial plane well outside the earth's atmosphere. The geomagnetic field lines passing through these points meet the earth's surface in latitudes corresponding roughly to the two auroral zones. Considerations of the mechanism for setting up this potential difference will be postponed to a later section, and meantime we will set no limits on its magnitude.

It is a well-known result of electrodynamics that the electrical conductivity, $\sigma$, of an ionized gas in the presence of a magnetic field, H, is not isotropic, but consists of three components which will be labeled $\sigma_0$, $\sigma_1$, and $\sigma_2$, (see, for instance, Baker and Martyn (7)). $\sigma_0$, the conductivity in the direction of the magnetic field, is unaffected by the presence of the field. $\sigma_1$, the transverse (Pedersen) conductivity, is reduced by a factor

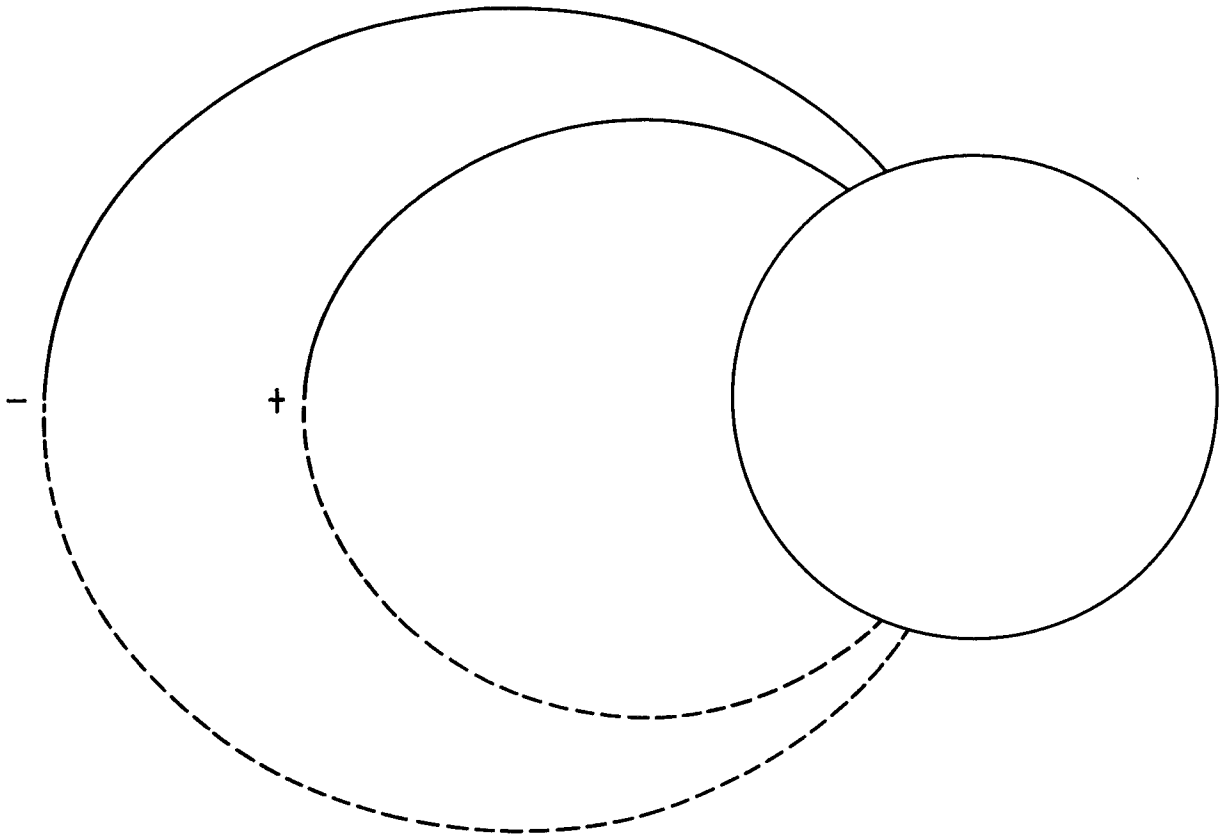


Fig. 1.

$\nu^2/(\omega_H^2+\nu^2)$, where $\nu$ is the collisional frequency and $\omega_H$ is the angular gyromagnetic frequency. The reduction is caused by the tendency of charged particles to spiral around magnetic lines of force, and the values to be adopted obviously depend on the type of particles under consideration. We will restrict ourselves to considering electrons. $\sigma_2$, the Hall conductivity, is in the direction of the vector product of the electric and magnetic fields, i.e., at right angles to both. In Fig. 1 it is directed into the paper, and its effect will not be considered for the moment, since it does not play any part in discharging the electric field.

In so-called interplanetary space, which is probably in reality an extension of the solar corona (see Chapman (8)), it is now fairly well agreed that the ion density is of the order of a few hundred per $cm^3$., and that the main constituent is probably hydrogen. This low value of ion density means that $\sigma_1$ is very much less than $\sigma_o$ at distances of the order of those in Fig. 1. Hence the potential difference will tend to discharge along the magnetic field lines into the ionosphere and in a meridional direction through the ionosphere at a height at which the transverse conductivity $\sigma_1$ becomes appreciable. Thus a steady current flow is visualized. Superimposed on this current is a drift of ionization in the east-west plane in the ionosphere caused by the Hall conductivity, $\sigma_2$. As mentioned above, this contributes nothing to the discharge, but may be the key factor governing the drift motions of auroral forms. It also provides a mechanism for transferring an electric field initially set up on the daylight side of the earth (towards the sun) round to the night side, where, as will be seen later, ionospheric conductivity conditions are such as to allow an auroral display to appear.

In order to simplify a quantitative estimate of the effects produced, let us imagine the potential drop along one magnetic field line alone, ignoring the transverse part of the ionospheric current, and the return current for the moment. It is reasonable to suppose that the electric field in a longitudinal direction at any point along this field line will be dependent on the local conductivity, by analogy with Ohm's Law. The situation is complicated, however, by the fact that an electric field increases the electron temperature in the gas, thereby decreasing the conductivity, which in turn tends to increase the electric field. We see intuitively from this that the electric field in the regions of low conductivity will become even greater at the expense of the field in more conductive regions. Chamberlain [5] has considered this effect, and has derived an expression for the current density, J, in a region of mean free path $\lambda$ when an electric field E is applied. His relation is

$$J = k N (Ee \lambda)^{2/3} \qquad (1)$$

where k is a constant, N is the local electron density and e is the electronic charge. This differs from the simple Ohm's Law relation

$$J = \sigma E \qquad (2)$$

because of the effect of the electric field on the collisional frequency. However, we can define an effective conductivity given by

$$\sigma = k N (e^2 \lambda^2 / E)^{1/3} \qquad (3)$$

Also, from (1), we have

$$E = \frac{1}{e\lambda}\left(\frac{J}{k\,N}\right)^{3/2} \tag{4}$$

Removing the constants,

$$E \propto \frac{J^{3/2}}{\lambda N^{3/2}} \tag{5}$$

Although the validity of this relation at very great heights is not certain, it has been used to give an order of magnitude for the field distribution along the entire magnetic field line.

The current density, J, has been taken as constant out to a distance of 1000 km. Beyond this the magnetic field lines begin to diverge appreciably, and since the total current flowing at all levels around the circuit must be constant in a steady state, the current density must begin to decrease. This has been taken account of by plotting dipole field lines originating in geomagnetic latitudes 66° and 69°, and measuring the separation between them. The relative current densities at different distances are then obtained from the fact that the product of current density and area remains constant.

An estimate of $\lambda$, the electron mean free path, as a function of height, was obtained by taking $\lambda = 1/\pi\, a_o^2\, N_o$, where $a_o$ is the classical radius of the electron and $N_o$ is the neutral atom density. Values of $N_o$ from 100 km. to 300 km., based on rocket measurements, have been tabulated by Mitra[9], and these were used in the calculation. Chapman[8] has calculated a value of $\lambda = 6.4 \times 10^7$ km. in the interplanetary gas in the neighborhood of the earth. A smooth curve was drawn through this point and the 100-300 km. points to obtain intermediate values of $\lambda$. The variation of $\lambda$ with height above the earth is shown in Fig. 2.

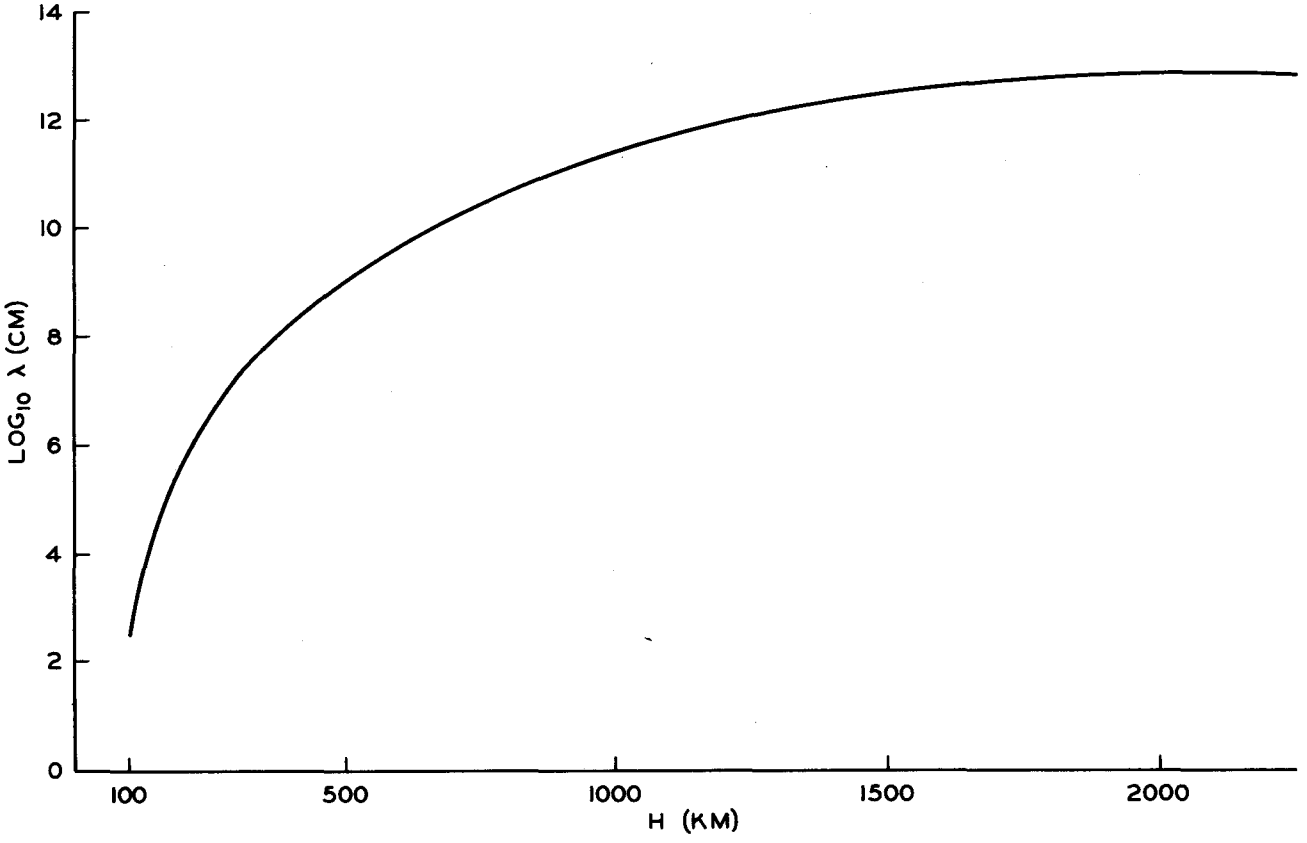


Fig. 2. Assumed variation with height of electron mean free path ( $\lambda$ ). The vertical scale is logarithmic.

The variation of N with height was estimated by adopting a value of $10^4$ per $cm^3$ at 100 km. under nighttime conditions (see Mitra (9)), rising to a maximum of $2.5 \times 10^5$ per $cm^3$ at 300 km., and then falling off with a scale-height of 300 km. to a constant value of 300 per $cm^3$ at great heights. The figure of 300 km. for the scale-height above the F2 maximum was adopted on the basis of some recent unpublished satellite measurements. The actual values to be adopted at great heights are not of much significance; only orders of magnitude are important at this stage. Fig. 3 shows the resultant variation of N with height.

For the purposes of estimating auroral effects, the important parameter is $E\lambda$, rather than E itself, since this gives a measure of the energy an electron can attain in the electric field, and hence the amount of excitation or ionization it can produce.

Fig. 4 shows the distribution of E with height, normalized to $E = 1$ at 100 km. It can be seen that the fall in field strength with height is very rapid, so that practically all of the initial potential difference appears at the base of the ionosphere. A rough integration of the curve in Fig. 4 shows that 98 per cent of the total potential drop along the field line appears between the levels of 100 km. and 120 km. in the ionosphere.

In Fig. 5 the product $E\lambda$ is plotted as a function of height. This curve shows some interesting features which form the basis of the theory of the initial stages of an auroral display. The curve has two maxima: the lower one, at the base of the E-region, is caused by the great concentration of electric field in this region shown in Fig. 4. The second maximum, at a height

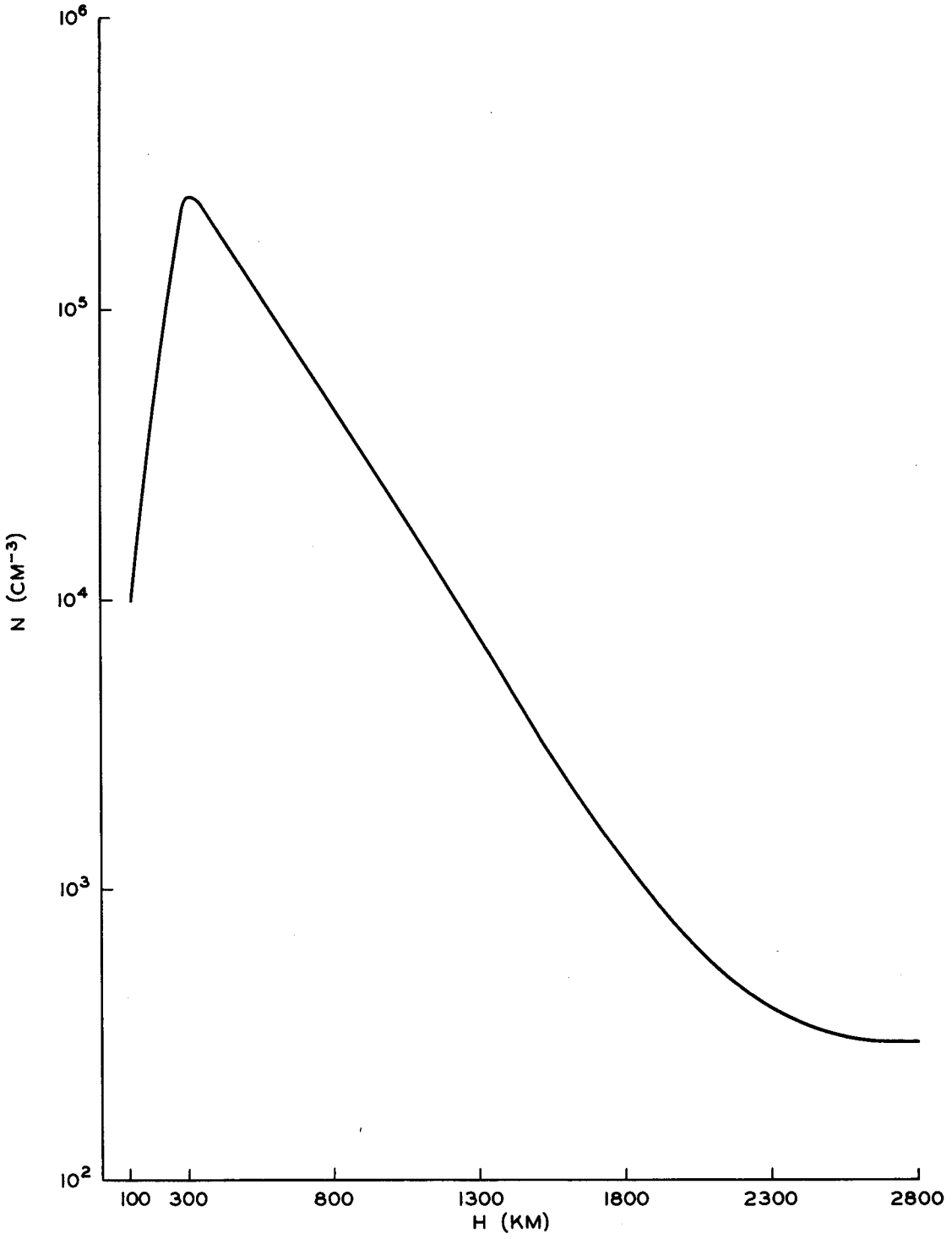


Fig. 3. Assumed variation with height of electron density (N). The vertical scale is logarithmic.

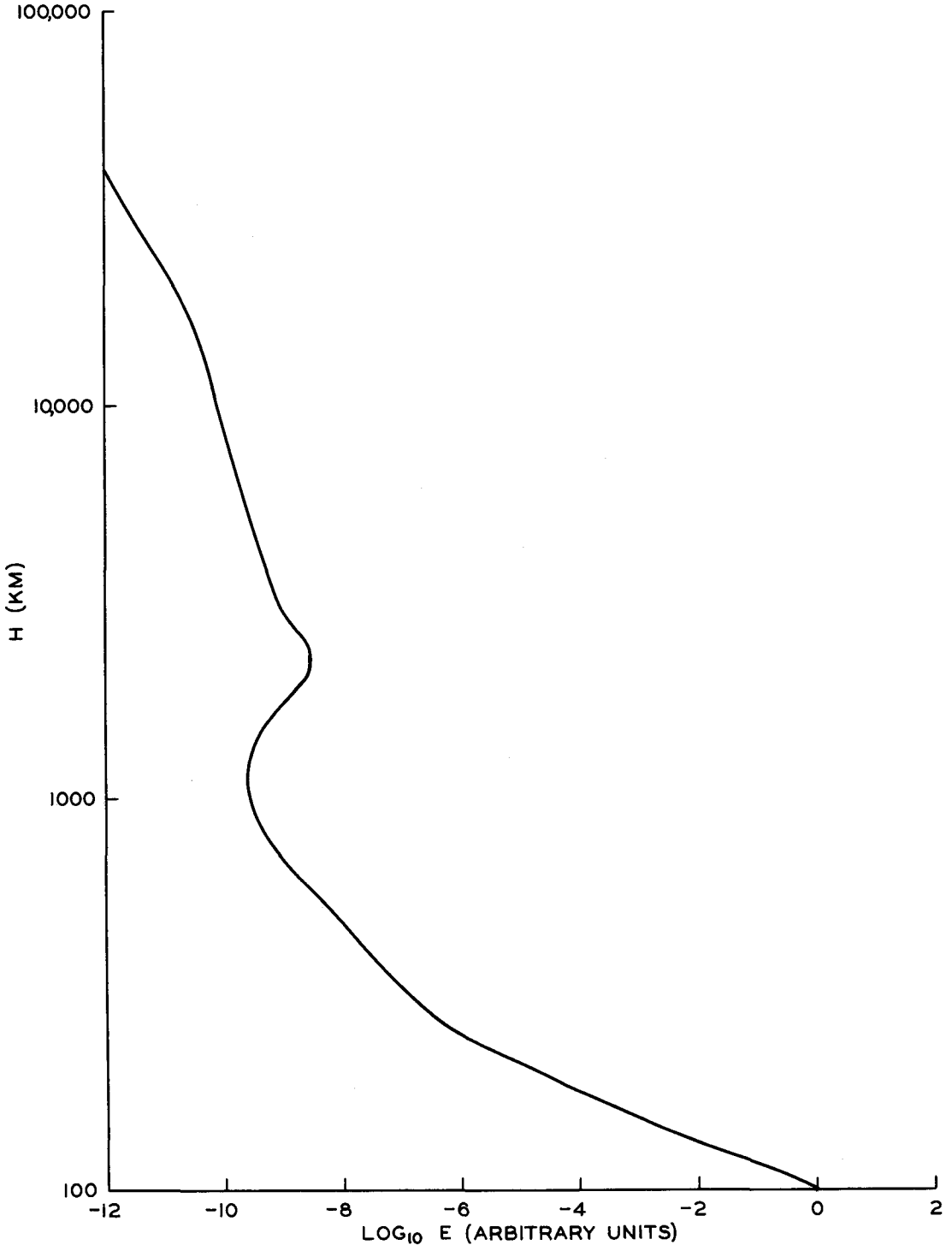


Fig. 4. Variation of electric field (E) with height. Both scales are logarithmic.

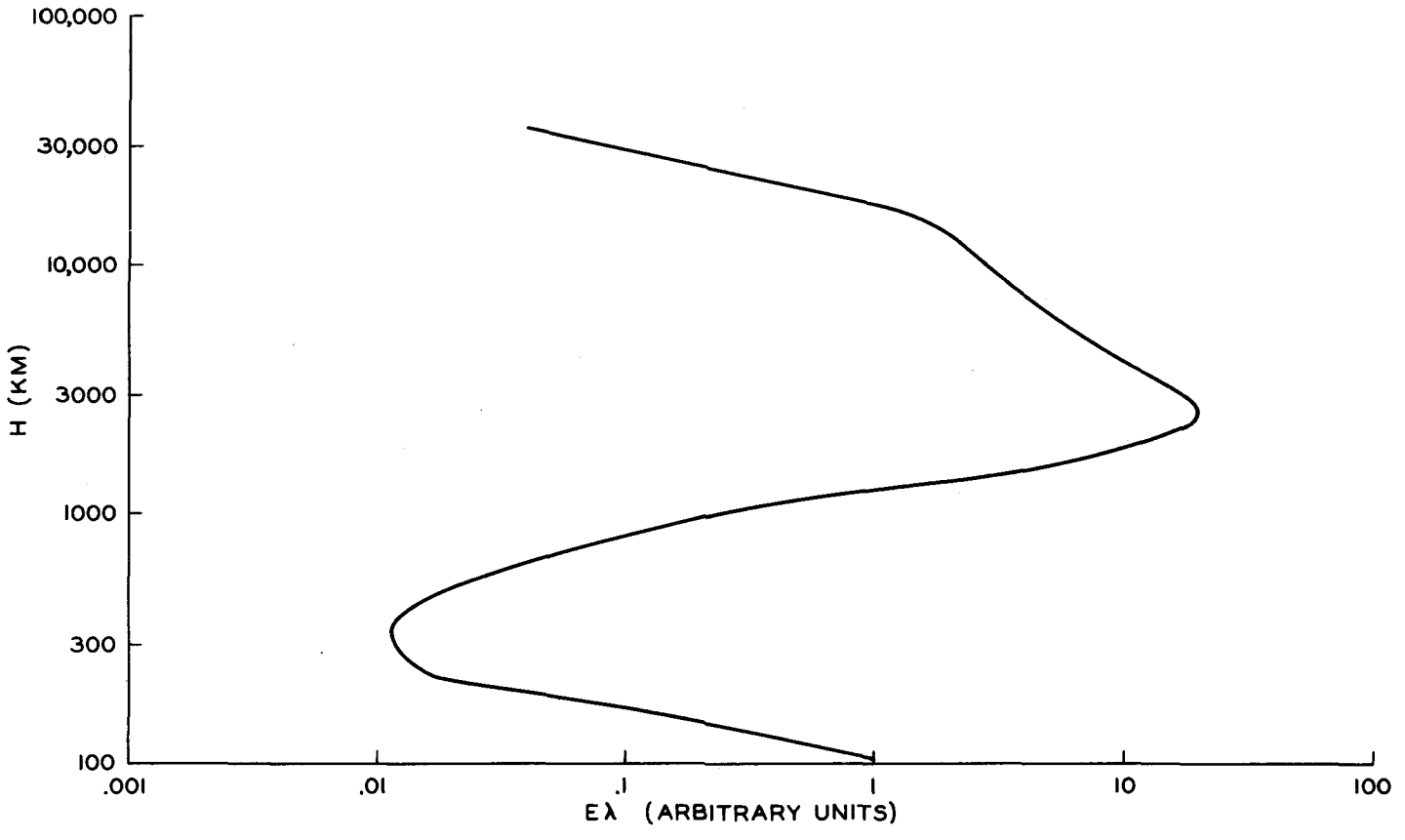


Fig. 5. Variation with height of E λ . Both scales are logarithmic

of about 2500 km., is caused by the very rapid increase in $\lambda$, which more than compensates for the decrease in field above the F-region. Above this level, $E\lambda$ again decreases, due to the divergence in the magnetic field lines and consequent decrease in current density, as mentioned above.

If we now imagine the electric field to increase steadily, excitation and emission of light will occur first at the upper maximum. At heights of the order of 2500 km., we are at the edge of the coronal gas, and the main constituent is probably hydrogen. The visible effect on the earth would be a faint diffuse glow of hydrogen emissions, and it is suggested that this is the explanation of the $H\beta$ emission which has been observed by Montalbetti[10] at Fort Churchill on occasional nights when no aurora was visible as such.

As the field continues to increase, excitation begins to take place at the 100 km. level, and light appears from a narrow band lying along a parallel of geomagnetic latitude, the appearance being that of a homogeneous arc. The hydrogen emissions will still be present, and will probably appear mainly to the south of the arc, due to the bending over of the field lines shown in Fig. 1. The arc will gradually increase vertically as the field increases, and eventually the field at 100 km. will become large enough to produce appreciable ionization. Immediately this happens, the conductivity at the E-region base will increase, leading to a lowering of the field, which in turn tends to stop the production of ionization. Hence the field in this region will be to some extent stabilized initially. This leads us to the conclusion that there should be very little extra ionization associated with the early stages of a homogeneous arc, and that most of the phenomena resulting from an increase in ionization, such as radio reflections and absorption, should occur after 'break-up',

when, as we shall see later, large amounts of ionization are produced. This result would be in contradiction to the predictions of any theory which explains a homogeneous arc as the direct result of incoming solar particles, which would produce large amounts of ionization.

Auroral Break-up

One of the most puzzling features of a typical auroral display is the so-called 'break-up'. After the initial quiet homogeneous arc stage, a rapid transition occurs to an active display of rays, draperies, pulsating and flaming aurora, accompanied by increased magnetic disturbance of the bay type, strong radio wave absorption, and often the production of X-rays. A qualitative explanation of this stage will be given here in terms of the potential drop discussed above.

So far we have considered the effects of applying a potential difference along the entire length of a magnetic field line. In the physical picture, however, there is also a current flow, and hence a potential difference, in a meridional direction (N-S) through the ionosphere. The magnitude of the field involved is difficult to estimate, since it depends on both the conductivity $\sigma_1$ and the current density $J_1$ which can flow. $\sigma_1$ is considerably less than $\sigma_0$ at most levels, which would tend to make the electric field in this direction much greater than in the vertical sections of the circuit. However, the meridional current flow takes place over a very much larger area, so that the current density is much less. This tends to reduce the necessary field, so that its order of magnitude might be comparable with that in the homogeneous arc.

If the field continues to increase beyond the point reached in the last section, eventually some of the meridional electrons will reach ionization energies. When this happens, the total 'resistance' of the circuit will rapidly decrease, leading to an increase in current flow. At the same time, the increased conductivity at the base of the E-region will cause a greater proportion of the total voltage drop to be applied to the vertical sections, so that ionization appears there in large quantities, tending to increase the current flow still more. At the 2500 km. level, it is suggested that the acceleration of the electrons may be violent enough to produce the X-rays which have been observed at lower levels by Winckler et al.[11]

It is beyond the scope of this paper to attempt any detailed explanation of the various auroral forms occurring during and after break-up, or the associated magnetic effects, and we will confine ourselves to pointing out the identification of the appearance of large amounts of ionization with the break-up. The auroral display is visualized as continuing until the initial potential drop has been neutralized.

Intense Auroral Displays

From time to time, very intense displays occur which are visible in latitudes far south of the normal auroral zone. This southward (or northward in the southern hemisphere) shift will be discussed later, and meantime we shall consider other striking features of these aurorae. The first characteristic is the abnormally red color of the entire display, which is caused by an enhancement of the red (6300 A) forbidden lines of atomic oxygen (9). This enhancement is thought to be a consequence of the abnormally great height at

which these displays occur, which is the second characteristic. Elvey[12] has reported measurements of a red homogeneous arc with a lower border at 335 km. on February 25, 1956. Measurements made during the intense red aurora of February 10, 1958 at the Geophysical Institute also indicated very great heights.

The third characteristic, which has not been noted before, is that, in the above two cases at least, the aurora was preceded at high latitudes by a period of very strong VHF radio wave absorption, of the kind which Reid and Collins[13] have called Type III. An excellent description of the absorption which occurred on February 23-25, 1956 has been given by Bailey[14]. This absorption indicates the presence of greatly increased amounts of ionization at low ionospheric levels, presumably due to incoming solar particles (low-energy cosmic rays). It has been found to decrease markedly after sunset, indicating negative-ion formation (15). If a similar increase in ionization took place at F-region heights, attachment would be a slow process, due to the low collisional frequency, so that the nighttime conductivity of the F-region would be abnormally high compared to that of the E-region. Under these conditions, the discharge might be expected to occur at higher levels than usual.

A similar argument can be applied to the observations of extremely high sunlit auroral rays reported by Störmer[2]. When sunrise occurs in the F-region, electrons are released, the conductivity increases, and more of the discharge can be carried by the higher regions.

The Diurnal and Seasonal Variations in Auroral Occurrence

During daytime, the total resistance around the circuit in Fig. 1 is very much less than at night, due to the greatly increased conductivity of the lower ionosphere. Consequently, the daytime values of $E\lambda$(which is proportional

to $N^{-3/2}$) would be much less than the nighttime values, so that little extra ionization would be produced. Hence the aurora as such should be a nocturnal phenomenon.

It may be pertinent to add that the 'circuit' as shown in Fig. 1 consists of two sections in parallel, one in each hemisphere. A rough idea of the effects of this on the seasonal variation of aurora can be gained from Fig. 6. Two points on the earth's surface have been selected which have the same magnetic longitude referred to the dip-pole, and equal and opposite magnetic latitudes; the E-region conductivity has been calculated as a function of local time at the northern station. For this purpose, the conductivity was taken as being proportional to the electron density, which in turn was taken as being proportional to $(\cos \chi)^{\frac{1}{2}}$, where $\chi$ is the sun's zenith angle. Fig. 6 shows the curves for two such points on January 15, April 15, July 15 and October 15. The points selected were Galena, Alaska (65°N, 157°W) and Campbell Island (52°S, 169°E); the appropriate curves are denoted by N and S respectively. The small residual nighttime conductivity has been neglected.

In January, we see that during the early evening, the northern path is effectively short-circuited by the highly conducting southern path, so that if aurora appears, it would not normally be expected to do so until late evening. The same effect might be expected to inhibit aurora in the southern region in July, though the effect will be smaller due to the difference in geographic latitude between the two points. During the equinoxes, the evening conductivities are more nearly equal at the two points, so that no preference will be given to either zone, and aurora will develop simultaneously at both when the electric field is sufficiently high.

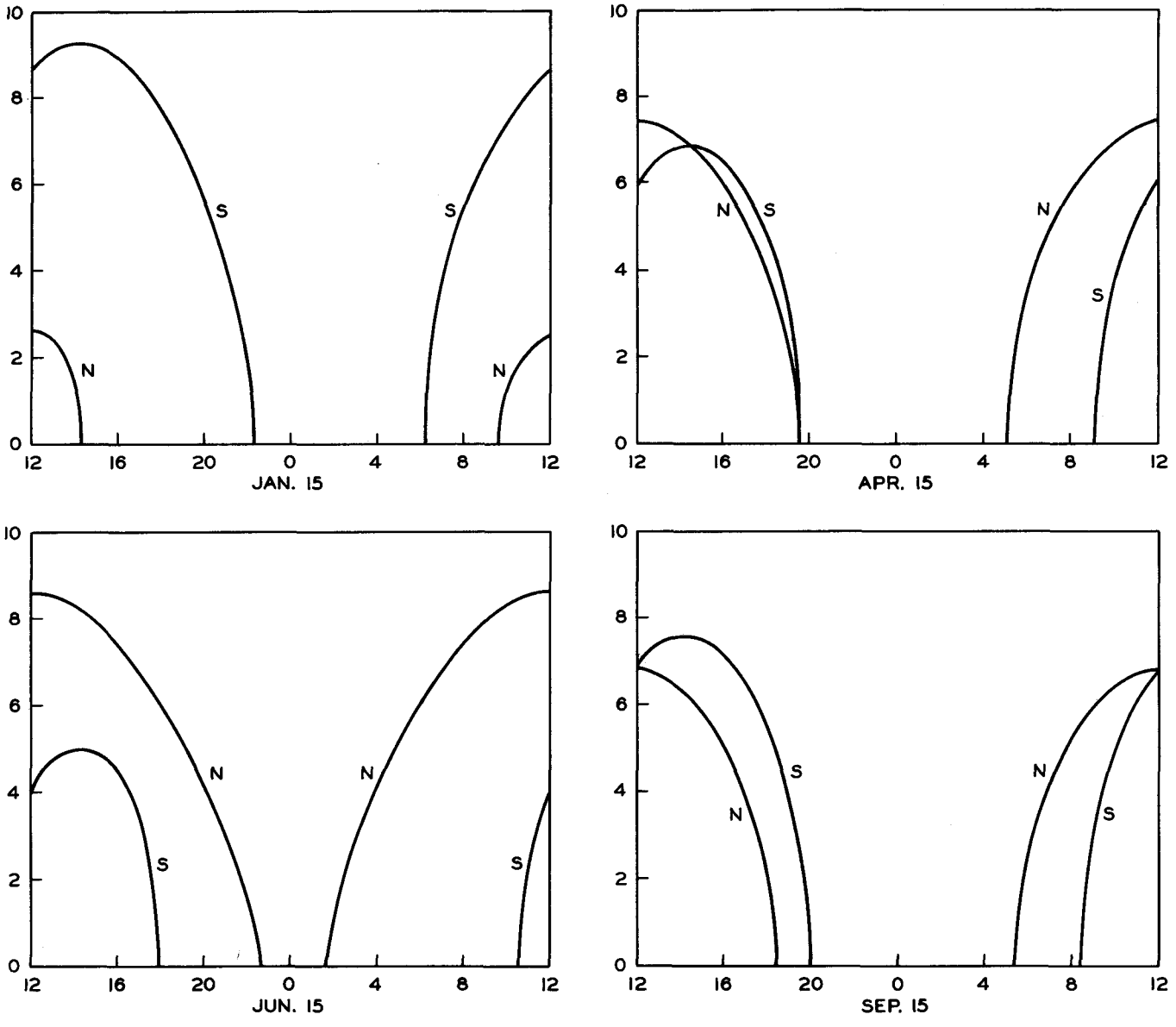


Fig. 6. Variation with time of E-region conductivity at two points at either end of a magnetic field line. Vertical scale is arbitrary; horizontal scale is local mean solar time at the northern point (N).

## The Origin of the Electric Field

The question of the origin of the electric field has been left to the last since any hypothesis must be highly speculative. As mentioned above, both Alfven and Martyn have proposed theories involving the development of electric fields as a consequence of the interaction between charged particles of solar origin and the geomagnetic field. Martyn's theory, which was an extension of the work of Chapman and Ferraro, made use of a hydrodynamical analogy to postulate the existence of a ring-current around the earth at a distance of several earth-radii, and found that a polarization field between the inner and outer surfaces of the ring was necessary for stability. Rough calculation shows that the total potential drop might be of the order of $10^6$ or $10^7$ volts. The formation of the ring has been questioned by many critics, and its existence has never been directly proved.

Recent rocket and satellite experiments, however, have shown the existence of a band of intense radiation at great heights above the earth's surface. The most likely explanation of this radiation appears to be that it consists of charged particles spiralling back and forth in the trapped orbits predicted by Störmer's theory (see Singer (16)). It seems reasonable to suppose that if the particles arriving from the sun consist, say, of protons and electrons of equal velocity, the electrons will tend to be trapped in orbits farther from the earth than the protons, by virtue of their smaller magnetic stiffness. This will produce a charge separation and an electrostatic field of the type shown in Fig. 1. Initially this field will be neutralized by a polarization of the ionized interplanetary medium, but if the incoming wave of particles is very dense, or continues for a long period, the density of trapped particles may exceed that of the medium, and the field will start to grow.

A simplified picture of events following a large solar flare might consist of four stages:

(1) Particles having a wide range of energies are emitted from the sun. The very high energy particles reach the earth in a short time and are occasionally observed at ground level as cosmic rays, as in the February 23, 1956 event. More often, their energy is not high enough for this, and they only penetrate a short distance into the atmosphere at high latitudes, producing Type III VHF radio wave absorption (see Reid and Collins (13); Leinbach and Reid (17)).

(2) Some of these particles are thrown into trapped orbits. The density of trapped particles gradually increases as particles of successively lower energies arrive.

(3) After the density has increased to a value greater than that of the neighboring interplanetary medium, an electric field starts to grow, which is conducted into the ionosphere via the magnetic field lines and produces an aurora. When the number of particles is unusually great, the trapped orbits will presumably expand in volume, the positive field line in Fig. 1 will be closer to the earth than usual, and the discharge will take place over a very great area of the earth's surface. This may be the cause of the great latitude extent of the aurorae which follow unusually violent solar eruptions.

(4) After new particles stop arriving, the field will gradually decay due to the current flow, and to the scattering of particles out of the trapped orbits.

## Conclusion

The foregoing material has been presented in the hope that it may prove to offer an interesting line of thought to auroral theorists.

The author is fully aware that the picture is far from complete, and that the lack of mathematical development may have tended to conceal inherent flaws in the process. However, at this stage, it is felt that a simple physical picture may be of more general interest than a detailed examination, which will no doubt follow. The lack of any investigation of the associated magnetic events can be criticized, and the author can only plead that others are much more fully qualified to examine these aspects of the theory.

## Acknowledgments

The author wishes to acknowledge his deep indebtedness to his colleagues at the Geophysical Institute, and particularly to Dr. C. T. Elvey for a great deal of encouragement and helpful discussion of the material.

References


(1) Chamberlain, J. W., Theories of the Aurora. 'Advances in Geophysics' (Academic Press, Ind.), Vol. 4 (1958).

(2) Störmer, C., 'The Polar Aurora'. Oxford University Press (1955).

(3) Lebedinski, A. I., The Ray and Arc Forms of the Aurora. Doklady Akad. Nauk SSSR, 86, 913 (1952).

(4) Alfven, H., 'Cosmical Electrodynamics'. Oxford University Press (1950).

(5) Chamberlain, J. W., Discharge Theory of Auroral Rays. 'The Airglow and Aurora' (A. Dalgarno and E. Armstrong, eds.) Pergamon Press (1956).

(6) Martyn, D. F., The Theory of Magnetic Storms and Auroras. Nature, 167, 92 (1951).

(7) Baker, W. G. and Martyn, D. F., Electric Currents in the Ionosphere. I. The Conductivity. Phil. Trans. Roy. Soc. London A, 246, 281 (1954).

(8) Chapman, S., Notes on the Solar Corona and the Terrestrial Ionosphere. Smithsonian Contribs. to Astrophys. 2, No. 1 (1957).

(9) Mitra, S. K., 'The Upper Atmosphere'. The Asiatic Society, Calcutta (1952).

(10) Montalbetti, R., Photoelectric Measurements of Hydrogen Emissions in Aurorae and Airglow. (In publication).

(11) Winckler, J. R., Peterson, L., Arnoldy, R. and Hoffman, R., X-rays from Visible Aurorae at Minneapolis. Phys. Rev. 110, 1221 (1958).

(12) Elvey, C. T., Problems of Auroral Morphology. Proc. Nat. Acad. Sci. 43, 63 (1957).

(13) Reid, G. C. and Collins, C., Observations of Abnormal VHF Radio Wave Absorption at Medium and High Latitudes. J. Atm. Terr. Phys. (In publication).

(14) Bailey, D. K., Disturbances in the Lower Ionosphere Observed at VHF Following the Solar Flare of 23 February 1956 with Particular Reference to Auroral-Zone Absorption. J. Geophys. Res. 62, 431 (1957).

(15) Chapman, S. and Little, C. G., The Non-Deviative Absorption of High-Frequency Radio Waves in Auroral Latitudes. J. Atm. Terr. Phys. 10, 20 (1957).

(16) Singer, S. F., 'Radiation Belt' and Trapped Cosmic-Ray Albedo. Phys. Rev. Letters, 1, 171 (1958).

(17) Leinbach, H. and Reid, G. C., (In publication).